PTO 05-1620

Japanese Kokai Patent Application
No. P2001-318792A

INTRINSIC REPRESENTATION EXTRACTION RULE GENERATING SYSTEM AND METHOD, RECORDING MEDIUM RECORDING PROCESSING PROGRAM THEREOF, AND INTRINSIC REPRESENTATION EXTRACTING DEVICE

Hideki Isozaki

UNITED STATES PATENT AND TRADEMARK OFFICE
WASHINGTON, D.C. JANUARY 2005
TRANSLATED BY THE RALPH MCELROY TRANSLATION COMPANY

JAPANESE PATENT OFFICE
PATENT JOURNAL (A)
KOKAI PATENT APPLICATION NO. P2001-318792A

Int. Cl.[7]:	G 06 F 9/44 17/28 17/30
Filing No.:	2000-137545
Filing Date:	May 10, 2000
Publication Date:	November 16, 2001
No. of Claims:	12 (Total of 11 pages)
Examination Request:	Not filed

INTRINSIC REPRESENTATION EXTRACTION RULE GENERATING SYSTEM AND METHOD, RECORDING MEDIUM RECORDING PROCESSING PROGRAM THEREOF, AND INTRINSIC REPRESENTATION EXTRACTING DEVICE

[Koyu hyogen chushutsu kisoku seisei shisutemu to hoho oyobi sono kyori puroguramu okirokushita kiroku baitai narabini koyou hyogen chushutsu sochi]

Inventor: Hideki Isozaki

Applicant: Nippon Telegraph & Telephone Corp.

[There are no amendments to this patent.]

Claims

/2*

1. An intrinsic representation extraction rule generating system characterized by the following facts: the intrinsic representation extraction rule generating system performs computer processing to generate the rule for use in extracting the intrinsic representations from a document on the basis of the document for training data in a storage device beforehand and a correct answer list that lists what is contained as the intrinsic representations (correct answer intrinsic

* [Numbers in the margin indicate pagination in the foreign text.]

representations) at what positions in the document for training for extracting what type of intrinsic representations; and it has the following means: a word type/character type attaching means, which reads said document for training from said storage device and divides it into words, attaches word type and structural character type to each word, generates a word row information that forms the intrinsic representations contained in said document for training and stores it in said storage device; a rule generating means, which reads the various correct answer intrinsic representations of said correct answer list from said storage device, compares them with the various word row information generated with said word type/character type attaching means, and generates the rule for extracting said correct answer intrinsic representations; a means for application of the rule for training, which reads said document for training and said rules from said storage device, applies said rules in said document for training, extracts the corresponding intrinsic representations (candidate intrinsic representations) and records them in said storage device; a rule evaluating means, which reads said candidate intrinsic representations and the correct answer intrinsic representations of said correct answer list from said storage device, compares them with each other, and computes the appropriateness of each rule used in extracting each candidate intrinsic representations on the basis of a prescribed computing sequence; a rule deleting means that deletes the rule with an appropriateness computed using said rule evaluating means lower than a prescribed appropriateness from said storage device; and a rule refining means that corrects the rule having the appropriateness computed using said rule evaluating means within a prescribed appropriateness range so as to increase its appropriateness and records the corrected rule in said storage device.

2. The intrinsic representation extraction rule generating system described in Claim 1 characterized by the fact that said rule generating means performs the following operation: when a word contained in the word row information read from said storage device is a numeral or a proper noun, or when the word is neither the word at the tail of said word row information nor any of the functional words including symbols, single kanji, tail connecting words, head connecting words, and particles, said word is converted to a variable, and a word row information containing variables is determined, and said rule is generated on the basis of said word row information containing variables and on the basis of said correct answer list.

3. An intrinsic representation extraction rule generating system characterized by the following facts: the intrinsic representation extraction rule generating system performs computer processing to generate the rule for use in extracting the intrinsic representations from a document on the basis of the document for training data in a storage device beforehand and a correct answer list that lists what is contained as the intrinsic representations (correct answer intrinsic representations) at what position in the document for training for extracting what type of intrinsic representations; and it has the following means: a word type/character type attaching means,

which reads said document for training from said storage device and divides it into words, attaches word type and structural character type to each word, generates word row information that forms the intrinsic representation contained in said document for training and stores it in said storage device; and a rule generating means that performs the following operation: said word row information is read from said storage device; when a word contained in said word row information read from said storage device is a numeral or a proper noun, or when the word is neither the word at the tail of said word row information nor any of the functional words including symbols, single kanji, tail connecting words, head connecting words, and particles, said word is converted to a variable, and word row information containing variables is determined, and said rule is generated on the basis of said word row information containing variables and on the basis of said correct answer list.

4. The intrinsic representation extraction rule generating system described in any of Claims 1-3 characterized by the fact that said rule generating means has a means that attaches to the generated rule a priority of the rule defined as the total number of rounds in which said intrinsic representation used in generating the rule appears in said correct answer list.

5. A type of intrinsic representation extracting device characterized by the following facts: the intrinsic representation extracting device has the intrinsic representation extraction rule generating system described in any of Claims 1-4, and it can extract the intrinsic representations contained in any document by means of computer processing on the basis of the rule generated with said intrinsic representation extraction rule generating system; in this intrinsic representation extracting device, there is a means that performs the following operation: when there is a partial overlap between plural extracted candidate intrinsic representations, the candidate intrinsic representation having an earlier description start position in said any document is extracted with priority; if they have the same description start position, the candidate intrinsic representation having a later description end position is taken as priority in extraction; also, there is a means that performs the following operation: when the plural extracted candidate intrinsic representations are the same, the candidate intrinsic representation having a higher priority attached beforehand to said rule used in extracting said candidate intrinsic representation is taken as the priority in extraction.

6. An intrinsic representation extraction rule generating method characterized by the following facts: in the intrinsic representation extraction rule generating method, computer processing is performed to generate the rule for use in extracting the intrinsic representations from a document on the basis of the document for training data in a storage device beforehand and a correct answer list that lists what is contained as the intrinsic representations (correct answer intrinsic representations) at what position in the document for training for extracting what type of intrinsic representations; and it has the following steps of operation: a first step in which

said document for training is read from said storage device and is divided into words, a second step in which the word type and structural character type are attached to each divided word to generate a word row information that forms the intrinsic representation contained in said document for training; a third step in which the various correct answer intrinsic representations of said correct answer list are read from said storage device and are compared with the various word row information generated in said second step to generate the rule for extracting said correct answer intrinsic representation; a fourth step in which said document for training and said rules are read from said storage device, said rules are applied in said document for training, and the corresponding intrinsic representation (candidate intrinsic representation) is extracted and recorded in said storage device; a fifth step in which said candidate intrinsic representation and said correct answer intrinsic representation of said correct answer list are read from said storage device and they are compared with each other, and the appropriateness of each rule used in extracting each candidate intrinsic representation is computed on the basis of a prescribed computing sequence; a sixth step in which the rule with an appropriateness computed using said rule evaluating means lower than a prescribed appropriateness is deleted from said storage device; and a seventh step in which the rule having the appropriateness computed using said rule evaluating means within a prescribed appropriateness range is corrected so as to increase its appropriateness, and the corrected rule is recorded in said storage device. /3

7. The intrinsic representation extraction rule generating method described in Claim 6 characterized by the fact that said third step has the following steps of operation: a step in which the following operation is performed: when a word contained in the word row information read from said storage device is a numeral or a proper noun, or when the word is neither the word at the tail of said word row information nor any of the functional words including symbols, single kanji, tail connecting words, head connecting words, and particles, said word is converted to a variable, and a word row information containing variables is determined, and a step in which said rule is generated on the basis of said word row information containing variables and on the basis of said correct answer list.

8. The intrinsic representation extraction rule generating method described in Claim 6 or 7 characterized by the following facts: said fourth step has a step in which the description position information of said candidate intrinsic representations in said document for training and the identification information of rule used in extracting said intrinsic representations are attached to said candidate intrinsic representations; said fifth step has the following steps: a step in which said candidate intrinsic representations and said correct answer list are read from said storage device and are compared with each other, and said extracted candidate intrinsic representations are classified to candidate intrinsic representations (intermediate candidate intrinsic representations) that are not in said correct answer list yet have their output suppressed by the

other correct answer intrinsic representations in said correct answer list, and candidate intrinsic representations (non-correct answer candidate intrinsic representations) that are not in said correct answer list and have their output not suppressed by the other correct answer intrinsic representations in said correct answer list, and a step in which for each rule used in extraction of the candidate intrinsic representations, the number of said correct answer intrinsic representations extracted with said rule and the number of said non-correct answer candidate intrinsic representations are counted; in said sixth step, the rule for which the number of said non-correct answer candidate intrinsic representations with respect to the number of said corrected answer candidate intrinsic representations is over a prescribed standard T1 is deleted from the rule group generated in said fourth step; in said seventh step, the rule for which the number of said non-correct answer candidate intrinsic representations with respect to the number of said corrected answer candidate intrinsic representations is lower than a prescribed standard T2 is corrected so that said number of the non-correct answer candidate intrinsic representations is reduced.

9. The intrinsic representation extraction rule generating method described in any of Claims 6-8 characterized by the following facts: in said fifth step, plural candidate intrinsic representations are read from said storage device with the same rule, they are classified into candidate intrinsic representations (correct answer candidate intrinsic representations) that are in agreement with said corrected answer intrinsic representations, candidate intrinsic representations (non-correct answer candidate intrinsic representations) that are not in agreement with said corrected answer intrinsic representations, candidate intrinsic representations (intermediate candidate intrinsic representations), and candidate intrinsic representations (intermediate candidate intrinsic representations) that are not in agreement with said correct answer intrinsic representation yet have their output suppressed with other said corrected answer candidate intrinsic representations, and computes said appropriateness of said corrected answer candidate intrinsic representations and non-correct answer candidate intrinsic representations on the basis of their numbers; in said seventh step, for each candidate intrinsic representation extracted by applying said rule (original rule) with said appropriateness in the prescribed appropriateness range, in said document for training, the words before and after it as well as the character types and word type of the words are determined, and on the basis of said words before and after the candidate intrinsic representation as well as the character type and word type of the word, a restricting condition, which ensures that said non-correct answer candidate intrinsic representation contained in each said candidate intrinsic representation is not extracted, is generated and added to said original rule.

10. An intrinsic representation extraction rule generating method characterized by the following facts: the intrinsic representation extraction rule generating method is adopted to

perform computer processing to generate the rule for use in extracting the intrinsic representations from a document on the basis of the document for training data in a storage device beforehand and a correct answer list that lists what are contained as the intrinsic representations (correct answer intrinsic representations) at what position in the document for training for extracting what type of intrinsic representations; and it has the following steps of operation: a first step in which said document for training is read from said storage device and is divided into words; a second step in which the word type and structural character type are attached to each word to generate a word row information that forms the intrinsic representations contained in said document for training, and it is recorded in said storage device; and a third step in which the following operation is performed: when a word contained in the word row information read from said storage device is a numeral or a proper noun, or when the word is neither the word at the tail of said word row information nor any of the functional words including symbols, single kanji, tail connecting words, head connecting words, and particles, said word is converted to a variable, and a word row information containing variables is determined, and said rule is generated on the basis of said word row information containing variables and on the basis of said correct answer list.

11. The intrinsic representation extraction rule generating method described in any of Claims 6-10 characterized by the fact that in said third step, a priority of the rule defined as the total number of rounds in which said intrinsic representation used in generating the rule appears in said correct answer list is attached to the generated rule.

12. A type of recording medium characterized by the following facts: the recording medium is for recording a program in a computer readable manner, with said program describing the processing of the method for generating the rule for use in extracting the intrinsic representations from a document on the basis of a document for training data in a storage device beforehand and a correct answer list that lists what is contained as the intrinsic representations (correct answer intrinsic representations) at what position in the document for training for extracting types of intrinsic representations.

Detailed explanation of the invention

[0001]

Technical field of the invention

The present invention pertains to a technology for extracting the intrinsic representation contained in a document by means of a computer. Especially, the present invention pertains to an intrinsic representation extraction rule generating system and its method that can be used preferably in generating the rule for extracting the intrinsic representations at a high efficiency, /4

as well as a type of recording medium for recording the processing program for said system and method, and a type of intrinsic representation extracting device.

[0002]

Prior art

In order to answer inquiries regarding the information contained in a large document, to make a summary of the document, to form a data base of the document or to visualize the document, it is necessary to extract the intrinsic representations, such as personal names, addresses, institution names, date/time, etc., from the document. In this case, one can make use of a computer to prepare a glossary that has the various intrinsic representations registered in it, and, by searching the glossary, one can perform extraction of the intrinsic representations from the document.

[0003]

However, the actual document may contain new words that are not included in the glossary prepared beforehand. Consequently, searching in the glossary only may not give a correct extraction result. In order to cope with this problem, there is the following technology: plural rules that can regulate the appearing pattern of the order of the intrinsic representation itself and the words before and after it are prepared manually beforehand; on the basis of the rules, computer processing is performed to extract the intrinsic representation from the document as the object.

[0004]

However, in this technology, the rules compete with each other and interact with each other. Consequently, the rules may not work as intended. As a result, the prepared rule has to be applied on certain training data prepared beforehand, and, on the basis of the result, if any error is observed, the rule is corrected. This operation is repeated in several rounds.

[0005]

However, as a result of correction of certain rule, the rules that used to operate normally may be affected, and erroneous answers may be given in many cases. Consequently, in order to have the plural rules all work as intended, a tremendous amount of labor is required.

[0006]

Even in the technology in which said rules for extracting the intrinsic representation are automatically generated using a computer, due to the competition and interaction between the

rules, a combination of said automatically generated rules is required to realize a good result, and this rule should be applied again, and the results are compared with the correct answer to be assessed. On the basis of the results, rules are added or deleted so as to get better results in repeated trial-and-error operation. This, however, requires a long computing time.

[0007]

Problems to be solved by the present invention

The problems to be solved are as follows: In the prior art, it is impossible to generate the rule for extracting the intrinsic representation contained in the document at a high precision, and in order to generate a better rule (rule for extracting the intrinsic representation), each time as the combination of the rules is corrected, it is applied on the practical document, and the result is compared with the correct answer to be graded, and trial-and-error is performed for the combination of the various rules. As a result, a huge computing time is needed, and this is undesirable.

[0008]

The purpose of the present invention is to solve the problems of the prior art by providing a type of an intrinsic representation extraction rule generating system and method that allow generation of high-precision intrinsic representation extracting rules easily in a short time and allow correct extraction of the desired intrinsic representations from a large document, as well as a recording medium that records the processing program and the intrinsic representation extracting device.

[0009]

Means to solve the problems

In order to realize the aforementioned purpose, in the intrinsic representation extraction rule generating system and method of the present invention, first of all, a document for training prepared beforehand is subjected to morphological analysis and is divided into words, and the information regarding the word type and structural character type, etc. is attached to each word. From the word row obtained in this way, the word row that forms the intrinsic representation is fetched, and by taking reference to the correct answer list prepared beforehand corresponding to the document for training, plural rules for extracting intrinsic representations are generated by means of empirical rules, minimum generalization, and other generalization means. Then, these rules are applied independently to the document for training, and the data regarding where the position in the document for training matches the rule are stored. These data become candidates of the intrinsic representation output from the system with respect to the document for training.

When plural rules are combined, from all of the candidates included in the data corresponding to said rules, the finally output candidate row is selected with a prescribed clear standard in consideration of the competitive relationship and the priority order. As a result, when a rule has a high frequency of non-correct answers or a very large proportion of said non-correct answers in the document for training, the rule is deleted. In this case, the word row before and after the correct answer site is compared with the word row before and after the non-correct answer site and a restriction is applied. As a result, it is possible to make a judgment on whether a rule with good results in the document for training is formed. Consequently, when the result is good, a rule with restriction applied on it is adopted.

[0010]

In addition, the intrinsic representation extracting device of the present invention has the intrinsic representation extraction rule generating system described above with the following features: it can extract the intrinsic representation in any document on the basis of the rule generated with said intrinsic representation extraction rule generating system. Also, when there is a partial overlap between plural extracted candidate intrinsic representations, the candidate intrinsic representation having an earlier description start position in said any document is extracted with priority; if they have the same description start position, the candidate intrinsic representation having a later description end position is taken as priority in extraction; also, when there are plural extracted candidate intrinsic representations with the same representation but of different types, the candidate intrinsic representation having a higher priority attached beforehand to said rule used in extracting said intrinsic representation is taken as the priority in extraction.

[0011]

Embodiments of the invention

In the following, the embodiments of the present invention will be explained in detail with reference to figures.

[0012]

Figure 1 is a block diagram illustrating an example of the constitution of the intrinsic representation extraction rule generating system of the present invention and the intrinsic representation extracting device having said intrinsic representation extraction rule generating system set in it. Figure 2 is a block diagram illustrating an example of the hardware constitution of the intrinsic representation extraction rule generating system and the intrinsic representation extracting device shown in Figure 1.

[0013] /5

In Figure 2, (21) represents a display device made of CRT (cathode ray tube), LCD (liquid crystal display), etc.; (22) represents an input device made of a keyboard, a mouse, etc.; (23) represents an external storage device made of HDD (hard disk drive) or the like; (24) represents an information processing device having (central processing unit) (24a), principal memory (24b), etc. and performing computer processing using the storage program system; (25) represents an optical disk made of CD-ROM (compact disk-read only memory) or DVD (digital video disk/digital versatile disk) or the like for recording the program and data pertaining to the present invention; (26) represents a driver for reading the program and data recorded on optical disk (25); and (27) represents a communication device made of LAN (local area network) card, modem, etc.

[0014]

After the program and data stored in optical disk (25) are installed in external storage device (23) via driver (26) by means of information processing device (24), they read from external storage device (23) to principal memory (24b), and are processed with CPU (24a). In said information processing device (24), there are both the intrinsic representation extraction rule generating system and the intrinsic representation extracting device having said intrinsic representation extraction rule generating system shown in Figure 1.

[0015]

In the intrinsic representation extracting device shown in Figure 1, document for training (1), correct answer list (2), intrinsic representation extraction rule group (5), improved intrinsic representation extraction rule group (5a), training data (7), novel document (11), and list (13) of the extracted intrinsic representations are stored in external storage memory (23) and principal memory (24b) shown in Figure 2. Also, morphological analysis/word type and character type attaching part (3), rule generating part (4), rule application part for training (6), rule evaluating part (8), rule deleting part (9), rule refining part (10), and rule application part for execution (12) are formed in information processing device (24) on the basis of the program stored in CD-ROM (25) shown in Figure 2.

[0016]

Said morphological analysis/word type and character type attaching part (3), rule generating part (4), rule application part for training (6), rule evaluating part (8), rule deleting

part (9), and rule refining part (10) form the intrinsic representation extraction rule generating system of the present invention.

[0017]

In morphological analysis/word type and character type attaching part (3), document for training (1) is divided into words, and information regarding the word type and the structural character type is attached to each word.

[0018]

In rule generating part (4), the word row obtained in the processing of morphological analysis/word type and character type attaching part (3) is compared with the data of the intrinsic representation to be extracted and given by correct answer list (2), and the word row that forms the intrinsic representation is fetched and generalized to generate a rule. The result is recorded as intrinsic representation extraction rule group (5) in external storage memory (23) in Figure 2.

[0019]

In rule application part for training (6), intrinsic representation extraction rule group (5) obtained as the result of generation of rule generating part (4) is applied in document for training (1). The result is recorded as data for training (7) in external storage device (23) in Figure 2.

[0020]

Rule evaluating part (8) evaluates the rules on the basis of data for training (7). On the basis of the evaluation result of rule evaluating part (8), rule deleting part (9) deletes the rule with poor results. Rule refining part (10) refines the rule so that the results become better.

[0021]

Rule application part for execution (12) applies the improved intrinsic representation extraction rule group (5) (improved intrinsic representation extraction rule group (5a)) on actual novel document (11) to obtain intrinsic representation list (13).

[0022]

For both rule application part for training (6) and rule application part for execution (12), the rule group is applied on the document to extract the intrinsic representation, and the processing contents are nearly the same. Consequently, it is possible to have both of them in the same device. Also, in rule application part for execution (12), there is no need to leave data for

training (7). However, it is necessary to perform selection of the final candidate. This is a point of difference.

[0023]

First, an explanation will be given regarding the operation of rule application part for execution (12), that is, the operation as an intrinsic representation extracting device using intrinsic representation extraction rule group (5) generated and improved with the intrinsic representation extraction rule generating system and improved intrinsic representation extraction rule group (5a).

[0024]

Rule application part for execution (12) applies improved intrinsic representation extraction rule group (5a) for novel document (11) for which the intrinsic representation is to be extracted, and it extracts the intrinsic representations contained in the document and outputs intrinsic representation list (13).

[0025]

For example, suppose there is new document (11) "In Tanaka Taro Prize Selecting Committee...", the intrinsic representations in this document include name candidates of "Tanaka", "Taro", "Tanaka Taro", an object name candidate of "Tanaka Taro Prize", as well as an institution candidate of "Tanaka Taro Prize Selecting Committee". Usually, it is demanded that among said candidates, the longest one, that is, "Tanaka Taro Prize Selecting Committee", be extracted and output as the intrinsic representation. In this case, the other candidates (intrinsic representations) of "Tanaka" and "Taro" overlapped with said intrinsic representation should not be output.

[0026]

The relationship among the candidates can be reduced to the competition relationship due to overlap and the suppression relationship due to the priority sequence of the various candidates. That is, because "Tanaka Taro Prize Selecting Committee" overlaps "Tanaka" and other candidates, they compete with each other. It is possible to consider that as the long candidate "Tanaka Taro Prize Selecting Committee" has a high priority, the other shorter candidates are suppressed.

[0027]

In this example, in rule application part for execution (12), on the basis of said consideration, first of all, all of the rules are adopted on the document, and a collection of all of the candidate intrinsic representations (including "Tanaka", "Taro", "Tanaka Taro", "Tanaka Taro Prize", "Tanaka Taro Prize Selecting Committee", etc.) is determined. Then, among said candidates, the longest candidate ("Tanaka Taro Prize Selecting Committee" among said candidates) of those having the same intrinsic representation ("Tanaka" in said candidates) is output.

[0028] /6

As one candidate is output, all of the other candidates ("Tanaka", "Tanaka Taro", "Tanaka Taro Prize") are deleted from the collection of the candidates. The aforementioned operation is performed repeatedly until the collection of candidates becomes empty. In this way, intrinsic representation list (13) is obtained.

[0029]

However, when only the length is taken into consideration, it is difficult to judge whether there are plural candidates having the same length by only performing judgment of selection from the various competing candidates. For example, "Whitehouse" may be taken as an address and an institution name. Consequently, the same character row "Whitehouse" is taken as both a candidate of address and a candidate of institution.

[0030]

In this case, for the two candidates, a priority order for extraction is set. For example, in consideration of the word before and after it, for "In the park near Whitehouse...", there is a high probability that it is an address. On the other hand, in "According to Whitehouse,...", it is quite possibly an institution name. Also, when the appearance frequency is taken into consideration, if there is only once when "Whitehouse" appears once in document for training (1), and there are 20 rounds in which it appears as an institution name, the possibility is high that it is judged as an institution name.

[0031]

In this example, a priority with said conditions taken into consideration is attached to each rule in improved intrinsic representation extraction rule group (5a).

[0032]

Rule application part for execution (12) combines such priority with said length of the intrinsic representation, and computes the priority order for each candidate. It is believed that there are various options in setting the priority order. However, as explained above, among those having the earliest start position and among those having the latest end position, it is clear that the candidate having the highest priority should be selected. That is, for the priority relationship of the candidates, the following definition is the basis.

[0033]

[1] If the start position of candidate A is earlier than that of candidate B (that is, a smaller numeral), candidate A has the priority.

[2] If the start position of candidate A is the same as that of candidate B, the candidate having the later end position (that is, a larger numeral) has the priority.

[3] When two candidates have the same start position and the same end position, the candidate having a larger priority u given according to the rule beforehand is taken as having the priority.

[0034]

In the intrinsic representation extraction rule generating system of this example, intrinsic representation extraction rule group (5) that allows easy processing with said rule application part for execution (12) and improved intrinsic representation extraction rule group (5a) are generated. In the following, an explanation will be given regarding the operations of the various parts that form the intrinsic representation extraction rule generating system pertaining to the generation processing of the rules with said priority relationship taken into consideration.

[0035]

First of all, in morphological analysis/word type and character type attaching part (3), the document is divided into words. The document, such as document for training (1) and new document (11), etc., having a typical morphological analysis function is divided into words. The word type and the type of the characters that form the word (structural character type information) are attached to each word to form a data structure, and a list is formed.

[0036]

For example, in the sentence "for Nakano, president of Tokyo Steel ....", results of morphological analysis indicate that "Tokyo" is a unique noun; "Steel" is an ordinate noun; "of" is a particle; "Nakano" is a unique noun; "president" is an ordinate noun; and "for" is a particle.

[0037]

Also, "Tokyo" is composed of plural kanji characters, and "NO [of]" is a Japanese character. Consequently, morphological analysis/word type and character type attaching part (3) outputs a list with the following data structure for said sentence. "(Tokyo, plural kanji characters, unique noun), (Steel, plural kanji characters, ordinate noun), (of, Japanese character, particle),...".

[0038]

On the other hand, correct answer list (2) lists the type of the intrinsic representation and the position in document for training (1). For example, correct answer list (2) prepared beforehand corresponding to document for training (1), "for Nakano, president of Tokyo Steel,...", is composed of the following data.

[0039]

0 3 東京製鉄① 組織名②
5 6 中野③ 人名④
20 23 3月9日⑤ 日付⑥
30 32 岡山県⑦ 地名⑧

Key:
1 Tokyo Steel
2 Institution name
3 Nakano
4 Person name
5 March 9
6 Date
7 Okayama
8 Place name

[0040]

In this list, in the first line, it is shown that "at the position from the 0^{th} character to the 3^{rd} character", "Tokyo Steel" of type of "institution name" is presented as an intrinsic representation. In the next line, "at the position from the 5^{th} character to the 6^{th} character", "Nakano" of type of "person name" is presented as an intrinsic representation. In correct answer list (2) of this example, the pair of numerals indicates the start position and end position of each intrinsic representation, and it gives a brief name indicating the position of the corresponding intrinsic representation.

[0041]

In rule generating part (4), said correct answer list (2) is compared with the word row output from morphological analysis/word type and character type attaching part (3), and it converts the intrinsic representations into variables. As a result, for example, the following rule for extracting the intrinsic representation is generated.

[0042]

angtag (3) ← <@(institution name, 21), word (_, plural kanji characters, unique noun), word (Steel, plural kanji characters, ordinary noun), >@ (institution name).

[0043]

According to this rule, the rule attaches number "21", and if there is any (in variable form) kanji unique noun ("word (_, plural kanji characters, unique noun)"), and the next word, "Steel" is an ordinary noun of plural kanji characters ("word (Steel, plural kanji characters, ordinary noun)"), these two words are taken as candidates of the intrinsic representation of "institution name".

[0044]

More generally, generation of said rule can be represented as follows. First of all, the intrinsic representation is composed of N+1 words $[(w_0, c_0, p_0), \cdots, (w_i, c_i, p_i), \cdots, (w_N, c_N, p_N)]$. Here, wi represents the word ("Steel", "Nakano", etc.), ci represents the structural character type ("plural kanji characters", "numeral", etc.), and pi represents the word type ("unique noun", "ordinary noun", etc.). /7

[0045]

In practice, the several surrounding words are also an important means in judging whether [the representation] is an intrinsic representation. Consequently, they are usually taken into consideration as well. However, in this specification, in order to simplify the discussion, only the words contained in the intrinsic representation are taken into consideration.

[0046]

Then, from the word row, minimum generalization or other existing generalization technology is used to generate the rule. However, in the present example, generation is performed in a simple way as follows.

[0047]

That is, the empirical rule to be explained later is applied on the specific word row $[(w_0, c_0, p_0), \cdots, (w_i, c_i, p_i), \cdots, (w_N, c_N, p_N)]$ that forms the intrinsic representation contained in document for training (1) to form a list $[(w_0', c_0', p_0'), \cdots, (w_i', c_i', p_i'), \cdots, (w_N', c_N', p_N')]$ containing variables, and the following rule is formed.

[0048]

$$\text{anytag}(u) \leftarrow \langle @(t+df, k), w_0 \\ rd(w_0', c_0', p_0'), \cdots, (w_i', c_i', p_i'), \\ \cdots, \text{word}(w_N', c_N', p_N'), >@(t-dt).$$

[0049]

Here, "t" indicates the type of the intrinsic representation (such as "institution name"). "+df" indicates how many characters should the start position of the intrinsic representation be shifted to the right, and it is a non-negative integer smaller than the number of characters of the initial word. Also, "-dt" indicates how many characters should the end position of the intrinsic representation be shifted to the left, and it is a non-negative integer smaller than the number of characters of the last word.

[0050]

For example, there is document for training (1) of "in Atsugi-shi,...". Although "Atsugi-shi" in it is a place name according to correct answer list (2), in the morphological analysis of morphological analysis/word type and character type attaching part (3), when it is divided to words of "Atsugi", "shi", "in", the word row that forms the intrinsic representation becomes "(Atsugi, plural kanji characters, unique noun), (shi, plural kanji characters, ordinary noun)", and the final one character ("in") is redundant. Here, in order to shift the end position by one character to the left, one has "dt=1". Also, because there is no shift for the start position, one has "df=0".

[0051]

Also, in said rule, "k" is a number attached to said rule, and "u" represents the priority of the rule.

[0052]

Data (wi', ci', pi') containing various variables is obtained as follows: corresponding to the data (wi, ci, pi) corresponding to the specific intrinsic representation contained in document for training (1), the following empirical rule is studied sequentially from the upper side, and the first matched one is adopted.

[0053]

[1] When "i" is "0" or "N", and the boundary of the intrinsic representation is contained ($df > 0$ or $dt > 0$), they are not formed as variables. In this case, in the rule, the original values of "df" and "dt" with respect to the original intrinsic representation are used as it is.

[2] For a numeral, "wi" is converted to a variable.

[3] For a unique noun, "wi" is converted to a variable.

[4] If the word is the last word of the list or a functional word, such as symbol, single kanji character, tail connecting word, head connecting word, particle, etc., no conversion to variable is performed.

[5] In other cases, "wi" is converted to variable.

[0054]

By applying the aforementioned processing for the various intrinsic representations, it is possible to automatically generate intrinsic representation extraction rule group (5).

[0055]

Also, as priority (u) of each rule, for example, the "total number of rounds" with which the intrinsic representation that becomes the origin of the rule appears in the correct answer list is adopted. As a result, it is possible to avoid the following problem that a rule with a smaller correct answer round number ("Whitehouse" as a place name in said example) suppresses a rule having a larger correct answer round number ("Whitehouse" as an institution name) without any justified reason.

[0056]

By applying the various rules (intrinsic representation extraction rule group (5)) obtained with said rule generating part (4) in the word row of document for training (1) in rule application part for training (6) to obtain training data (7). That is, in rule application part for training (6), from the head to the tail of document for training (1), the positions where the rules match are

studied sequentially. When matched, it is taken as a candidate and is added to training data (7). This operation is repeated.

[0057]

For training data (7), more specifically, comparison is performed for the competition relationship and suppressing relationship between the various candidates, and the data of rule number (k), matched position, type of the intrinsic representation (t), etc. are recorded such that the final output can be obtained.

[0058]

The processing with said rule application part for training (6) is performed for all of the rules of intrinsic representation extraction rule group (5) to form training data (7).

[0059]

Also, by means of a bottom-up type text analysis scheme, it is possible to simultaneously obtain plural rule application results at a high efficiency.

[0060]

Rule evaluating part (8) reads training data (7) prepared in the above, and makes grading for the result of each rule. Various standards may be adopted as the specification for grading. A simple way is to make use of the evaluation by means of the number of rounds and proportion of the non-correct answer. However, more strictly speaking, the number of rounds of non-correct answer for each rule depends on the rules combined with it. Consequently, when the specific rule to be adopted has not yet been decided, it is impossible to get a correct numeral. In this case, records of the rules (R) are classified as follows for consideration.

[0061]

(O) The candidate obtained by matching with the intrinsic representation as the base of rule R, that is, the candidate surely becomes correct answer if not suppressed with other candidate (correct answer candidate intrinsic representation).

(Δ) The other competing intrinsic representation is registered in correct answer list (2), and the candidate is suppressed by it. That is, if the intrinsic representation is not a correct answer, the output is suppressed, so that in the rule group with a high precision, the candidate has a high possibility without decrease in the result (intermediate candidate intrinsic representation).

(x) The others. That is, because there is no suppressed correct answer intrinsic

representation, in the high-precision rule group, there is a high possibility that a wrong candidate is output and the result decreases (non-correct answer candidate intrinsic representation).

[0062]

In rule evaluating part (8), the number of rounds is counted for each of "O", "Δ" and "x" for each rule, and the number of rounds of "x" is adopted as the number of rounds of non-correct answer, and the number of rounds of "O" is adopted as the number of the rounds of the correct answer. Also, if all of "Δ" are taken as non-correct answer, the rule that extracts "Tanaka" or other short intrinsic representation becomes unfavorable. Consequently, this should be avoided. For this purpose, in rule evaluating part (8), the following method is adopted to count the rounds of the non-correct answer.

[0063]

That is, rule evaluating part (8) sequentially reads training data (7) from the former side, and rule R is applied at position L of document for training (1). The type of the intrinsic representation attached with rule R (classification of place name, person name, etc.) is T; the pair of type T and location L is not contained as a correct answer in correct answer list (2); and, in addition, the intrinsic representation of the correct answer either is not present at the position overlapped with location L, or, although it is present, if the candidate according to rule R is prior to the candidate corresponding to the correct answer, the number of rounds of the non-correct answer of rule R is increased by one. This operation is performed repeatedly until end of training data (7).

[0064]

Rule evaluating part (8) counts the numbers of "O", "Δ", "x" of each rule. With respect to this result, rule deleting part (9) and rule refining part (10) apply correction on intrinsic representation extraction rule group (5).

[0065]

In the rules of intrinsic representation extraction rule group (5), for example, rule deleting part (9) deletes the rules which have the number of "x" larger than that of "O". Rule refining part (10) performs the following operation: in the rules of intrinsic representation extraction rule group (5), for example, a restriction information pertaining to the words before and after it is added to the rules which have the number of "x" in the result larger than half the number of "O", so as to improve the results of the rules.

[0066]

For example, suppose two words before the intrinsic representation and two words after the intrinsic representation are included for consideration, in each intrinsic representation extracted with said rule and classified by evaluation to "O" or "x", the word list of $[(w_{-2}, c_{-2}, p_{-2}), (w_{-1}, c_{-1}, p_{-1}), (w_0, c_0, p_0), \cdots, (w_{N+1}, c_{N+1}, p_{N+1}), (w_{N+2}, c_{N+2}, p_{N+2}),]$ is considered. In this case, for each intrinsic representation, the characteristic list of $(w_{-2}, c_{-2}, p_{-2}, w_{-1}, c_{-1}, p_{-1}, w_{N+1}, c_{N+1}, p_{N+1}, w_{N+2}, c_{N+2}, p_{N+2})$ is considered. Suppose the positive case is for the intrinsic representation classified as "O", and the negative case is for the intrinsic representation classified as "x", it is a typical topic of inductive learning, and the existing machine learning scheme can be used as is.

[0067]

For example, by means of learning using a determining tree, among the several words before and after the [intrinsic representation], it is possible to determine the value of what property of what word to be left, while the remainder is to be converted to variables. As a special example, suppose "10" intrinsic representations classified to "x" are extracted, and, among them, "8" intrinsic representations have "wx" specified as the preceding word (w-1), as shown below, a restrictive condition $\{w_{-1}' \neq w_X\}$ is applied on the original rule, and restriction is made such that the intrinsic representation having "wx" is not extracted as the preceding word (w-1).

[0068]

$$anytag(u) \leftarrow word(w_{-1}', c_{-1}', p_{-1}'), <@(t+df, k), word(w_0', c_0', p_0'), \cdots, (w_i', c_i', p_i'), \cdots, word(w_N', c_N', p_N'), >@(t-dt), \{w_{-1}' \neq w_X\}.$$

[0069]

For the rule obtained in this way, there is a strong restriction from the original rule. Consequently, matching takes place only for the portion identical to the portion that matches the original rule. Consequently, even when not adopted on the entirety of document for training (1), as long as it is applied only for the portion matched with the original rule left in training data (7), the results of the new rule can be understood.

[0070]

In this example, improvement of the rule is performed almost independent from other rules. As explained above, rules with better results (improved intrinsic representation extraction rule group (5a)) are generated from the original rules (intrinsic representation extraction rule group (5)).

[0071]

Figure 3 is a flow chart illustrating an example of the processing process of the intrinsic representation extraction rule generating method pertaining to the present invention.

[0072]

In this example, in the intrinsic representation extraction rule generating system shown in Figure 1, the various processing operations of morphological analysis/word type and character type attaching part (3), rule generating part (4), rule application part for training (6), and rule evaluating part (8) are shown. First of all, in morphological analysis/word type and character type attaching part (3), document for training (1) is subject to morphological analysis, and it is divided into words (step (301)), and the information of the word type and character type, etc. is attached to each word (step (302)).

[0073]

Then, in rule generating part (4), the intrinsic representation of correct answer list (2) and the word row composed of the words near it are extracted (step (303)), the empirical rule or the like is applied on the correct answer word row to generate extracting rules (step (304)), and they are recorded as intrinsic representation extraction rule group (5)).

[0074]

In rule application part for training (6), the extracting rules generated in this way are applied to document for training (1), and the intrinsic representation obtained as a result is recorded as a candidate (step (305)).

[0075]

In addition, in rule evaluating part (8), the correct answer degrees (O, Δ, x) of the intrinsic representations extracted with the various extracting rules are determined and classified. On the basis of said operation, the appropriateness of each extracting rule is assessed (step (306)).

[0076]

As the result of the grading, the rule group with a poor result that makes it not correctable (with a low appropriateness) is deleted in rule deleting part (9) (step (307)). Also, in rule refining part (10), said correction is applied on the rule group having a higher appropriateness by correction to form new rules (step (308)), and they are recorded as improved intrinsic representation extraction rule group (5a). By performing the processing from step (305) repeatedly, it is possible to generate a rule group with better results.

[0077]

Figure 4 is a flow chart illustrating an example of processing operation of the intrinsic representation extracting device shown in Figure 1. In this example, in the intrinsic representation extracting device shown in Figure 1, the processing operation for new document (11) is shown. First of all, in morphological analysis/word type and character type attaching part (3), new document (11) is subjected to morphological analysis and it is divided into words (step (401)), and the influence of the word type and character type, etc. is attached to each word list (step (402)). /9

[0078]

Then, in rule application part for execution (12), in each word list, the various extraction rules of improved intrinsic representation extraction rule group (5) are applied, and the various intrinsic representations are taken as candidates for list-up (step (403)), and for all of the candidates, the following priority control processing is performed (step (404)). That is, the candidate with the highest priority in the candidates is output (step (405)), and the candidates that compete with said output candidate are deleted (step (406)).

[0079]

In the aforementioned intrinsic representation extraction rule generating system and method explained with reference to Figures 1-4, first of all, morphological analysis is performed for document for training (1) prepared beforehand so that it is divided into words; the influence of the word type and structural character type, etc. is attached to each word; from the obtained words, the word row that forms the intrinsic representation is fetched; and, by means of the empirical rule, minimum generalization, or other generalizing means with reference to correct answer list (2) prepared corresponding to document for training (1) beforehand, plural intrinsic representation extracting rules are generated.

[0080]

Then, the extracting rules are independently applied on document for training (1), and data indicating the position of document for training (1) where the rules match is prepared. This data represents the candidates of the intrinsic representation output from the system with respect to document for training (1).

[0081]

Then, when plural rules are combined, from all of the candidates that enter the records corresponding to the rules, the row of candidates to be finally output are selected with a prescribed clear standard in consideration of the competition relationship and the priority order. As a result, the rule that has a very high frequency of non-correct answers or a very high proportion of the non-correct answers in document for training (1) is deleted. It is known that the rule is a correct answer at a certain position of the document for training, and it is a non-correct answer at certain other positions of the document for training. By applying a restriction by comparing the word row before and after the correct answer site with that before and after the non-correct answer site, it is possible to judge whether a rule that has a good result in the document for training has been formed. Consequently, when the results are good, the rule with restriction on it is applied.

[0082]

In this example, when a document for training containing the intrinsic representations and a correct answer list that lists what type of intrinsic representation in what position in the document are given, the system can generate the intrinsic representation extracting rules on the basis of the correct answer, there is no need to write the extracting rules, this saving a great deal of labor.

[0083]

In addition, evaluation is performed for the various rules output with respect to document for training (1) prepared beforehand. Then, the evaluation value is determined for the various combinations of plural rules by means of simple computation from the evaluation values of the various individual rules. As a result, it is possible to shorten the processing time needed for trial-and-error performed during the process of determination of the combination of rules with good results. Also, improvement of the intrinsic representation extracting rules is performed almost independent from other rules, and the precision can be improved easily.

[0084]

Also, in the intrinsic representation extracting device of this example, the rules generated and improved on the basis of the document for training and the correct answer list are applied on new document (11), and the intrinsic representations are automatically extracted from said new document (11). At the same time, if the extracted plural intrinsic representations are partially overlapped with each other, the intrinsic representation having an earlier description start position in said document is extracted with priority; if they have the same description start position, the intrinsic representation having a later description end position is taken as priority in extraction; also, when there are plural types of intrinsic representations having the same representation, the intrinsic representation having a larger priority attached beforehand to said rule used in extracting said intrinsic representation is taken as the priority in extraction. As a result, it is possible to perform extraction limited only to the appropriate intrinsic representation.

[0085]

The present invention is not limited to the example explained with reference to Figures 1-4. Various modifications can be made as long as the gist is observed. For example, in this example, when a restriction is attached to the rules, the restriction is set on the basis of the words before and after the candidate intrinsic representation in the document for training. However, it is also possible to set a restriction pertaining to the character type of the word (kanji character, Japanese character, ...) and word type (noun, verb, ...), etc.

[0086]

Also, in this example, optical disk (25) is used as the recording medium. However, one may also adopt FD as the recording medium. In addition, as far as installing of the program is concerned, it is also possible to go through communication device (27) to download the program via a network and then install it.

[0087]

Effect of the invention

According to the present invention, the rules for extracting the intrinsic representations are automatically generated on the basis of the document for training prepared beforehand and the correct answer list that lists what type of intrinsic representations are included at what positions in the document. Consequently, there is no need to write down the labor-intensive extracting rules. In addition, by comparing the result of application of the automatically generated rules on the document for training with the correct answer list, it is possible to determine the appropriateness of each rule and to determine the appropriateness of a combination

of various rules on the basis of the appropriateness of each rule. Consequently, improvement of the intrinsic representation extracting rule can be performed almost independent from the other rules, it is possible to improve the precision easily, and it is possible to realize a high-performance intrinsic representation extracting device.

Brief description of the figures

Figure 1 is a block diagram illustrating an example of the constitution of the intrinsic representation extraction rule generating system and the intrinsic representation extracting device having said intrinsic representation extraction rule generating system set in it in the present invention.

Figure 2 is a block diagram illustrating an example of the constitution of the hardware of the intrinsic representation extraction rule generating system and intrinsic representation extraction rule generating device shown in Figure 1.

Figure 3 is a flow chart illustrating an example of the processing process of the intrinsic representation extraction rule generating method in the present invention. /10

Figure 4 is a flow chart illustrating an example of the processing operation of the intrinsic representation extraction rule generating device shown in Figure 1.

Brief description of the reference numbers

1	Document for training
2	Correct answer list
3	Morphological analysis/word type and character type attaching part
4	Rule generating part
5	Intrinsic representation extraction rule group
5a	Improved intrinsic representation extraction rule group
6	Rule application part for training
7	Training data
8	Rule evaluating part
9	Rule deleting part
10	Rule refining part
11	New document
12	Rule application part for execution
13	Extracted intrinsic representation list
21	Display device
22	Input device
23	External storage device

24 Information processing device
24a CPU
24b Principal memory
25 Optical disk
26 Driver
27 Communication device

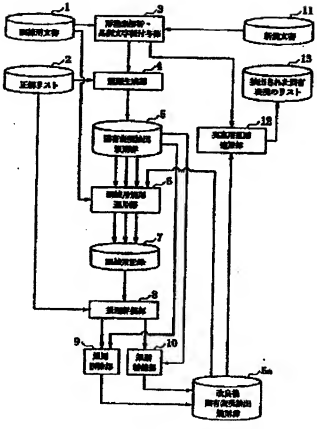


Figure 1

Key: 1 Document for training
2 Correct answer list
3 Morphological analysis/word type and character type attaching part
4 Rule generating part
5 Intrinsic representation extraction rule group
5a Improved intrinsic representation extraction rule group
6 Rule application part for training
7 Training data
8 Rule evaluating part

9 Rule deleting part
10 Rule refining part
11 New document
12 Rule application part for execution
13 Extracted intrinsic representation list

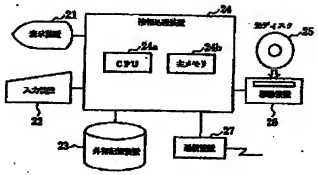


Figure 2

Key: 21 Display device
22 Input device
23 External storage device
24 Information processing device
24b Principal memory
25 Optical disk
26 Driver
27 Communication device

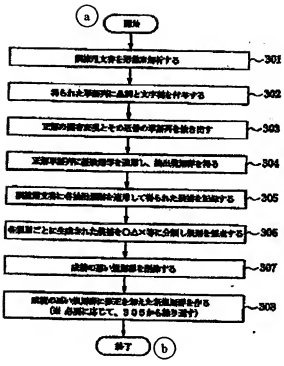


Figure 3

Key:
- a START
- b END
- 301 Morphological analysis of document for training
- 302 Attachment of word type and character type to obtained word row
- 303 Extraction of correct answer intrinsic representation and the word row before and after it
- 304 Application of empirical rule or the like on the correct answer word row to obtain extraction rule group
- 305 Recording of candidates obtained by applying various extracting rules on the document for training
- 306 Classification of candidates generated for each rule to O, Δ, X, etc., and assessment of the rule
- 307 Deleting of the rule group with poor results
- 308 Preparation of new rule group after correction of the rule group with poor results (* as needed, repeating from step (305))

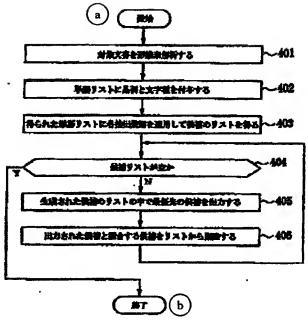


Figure 4

Key: a START
b END
401 Morphological analysis of the object document
402 Attachment of word type and character type to the word list
403 Application of each extracting rule on the obtained word list to obtain the candidate list
404 Is the candidate list empty?
405 Output of the candidate with the lowest priority in the generated candidate list
406 Deletion of the candidates that compete with the output candidate from the list

(19)日本国特許庁（ＪＰ）　(12)公開特許公報(A)　(11)特許出願公開番号
特開2001－318792
（P2001－318792A）
(43)公開日　平成13年11月16日(2001.11.16)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'(参考)
Ｇ０６Ｆ　9/44	５８０	Ｇ０６Ｆ　9/44	５８０Ｐ	５Ｂ０７５
17/28		17/28	Ｚ	５Ｂ０９１
17/30	１７０	17/30	１７０Ａ	
	１８０		１８０Ａ	
	２１０		２１０Ｚ	

審査請求　未請求　請求項の数12　ＯＬ　（全 11 頁）

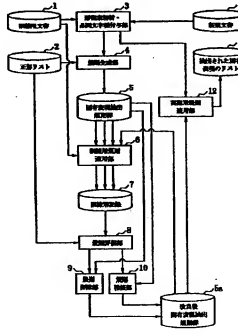
(21)出願番号　特願2000－137545(P2000－137545)
(22)出願日　平成12年５月10日(2000.5.10)

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(72)発明者　磯▲崎▼　秀樹
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(74)代理人　100077274
弁理士　磯村　雅俊　（外１名）

Ｆターム(参考)　５Ｂ０７５　ＮＤ０３　ＮＫ３２　ＮＫ３９
５Ｂ０９１　ＡＡ１５

(54)【発明の名称】　固有表現抽出規則生成システムと方法およびその処理プログラムを記録した記録媒体ならびに固有表現抽出装置

(57)【要約】
【課題】　従来の技術では、高精度な固有表現抽出規則の生成を容易に短時間で行うこと、および、膨大な文書から所望の固有表現を正確に抽出することができない。
【解決手段】　まず、形態素解析・品詞文字種付与部３により、予め用意された訓練用文書１から各単語を抽出し、次に、規則生成部４により、各単語と訓練用文書１に対応して予め用意された正解リスト２とに基づき固有表現抽出用の規則（ルール）を生成する。そして、訓練用規則適用部６により、これらの規則をそれぞれ独立に訓練用文書１に適用して、各規則毎の固有表現抽出結果を求め、規則評価部８により、その適用結果で得られた固有表現と正解リストとを比較して、各規則の適正度を比較し、その結果に基づき、規則削除部９と規則精錬部１０により、適宜、規則の削除と修正を行う。



1

【特許請求の範囲】

【請求項1】　文書から固有表現を抽出するのに用いる規則を、予め記憶装置に記録された訓練用文書と、該訓練用文書の中のどの位置にどのような種類の固有表現が抽出されるべき固有表現（正解固有表現）として含まれているかを列挙した正解リストとに基づきコンピュータ処理して生成する固有表現抽出規則生成システムであって、上記訓練用文書を上記記憶装置から読み出して単語分割し、各単語に品詞名と構成文字種を付加して、上記訓練用文書に含まれる固有表現を構成する単語列情報を生成して上記記憶装置に記録する品詞文字種付与手段と、上記正解リストの各正解固有表現を上記記憶装置から読み出して上記品詞文字種付与手段で生成された各単語列情報と比較し、当該正解固有表現を抽出するための規則を生成して上記記憶装置に記録する規則生成手段と、上記記憶装置から上記訓練用文書と上記規則とを読み出して各規則を上記訓練用文書に適用し、対応する固有表現（候補固有表現）を抽出して上記記憶装置に記録する訓練用規則適用手段と、上記記憶装置から上記候補固有表現と上記正解リストの正解固有表現とを読み出して比較し、予め定められた算出手順に基づき、各候補固有表現の抽出に用いた各規則の適正度を算出する規則評価手段と、該規則評価手段で算出した適正度が予め定められた適正度より低い規則を上記記憶装置から削除する規則削除手段と、上記規則評価手段で算出した適正度が予め定められた適正度範囲の規則を、当該適正度が高くなるように修正して上記記憶装置に記録する規則精錬手段とを有することを特徴とする固有表現抽出規則生成システム。

【請求項2】　請求項1に記載の固有表現抽出規則生成システムにおいて、上記規則生成手段は、上記記憶装置から読み出した単語列情報に含まれる単語が、数字か固有名詞の場合もしくは当該単語列情報の末尾の単語が記号・単漢字・接尾語・接頭語・助詞を含む機能語のいずれでもない場合には該単語を変数化して、変数を含む単語列情報を求め、該変数を含む単語列情報と、上記記憶装置から読み出した上記正解リストとに基づき上記規則を生成する手段を有することを特徴とする固有表現抽出規則生成システム。

【請求項3】　文書から固有表現を抽出するのに用いる規則を、予め記憶装置に記録された訓練用文書と該訓練用文書の中のどの位置にどのような種類の固有表現が抽出されるべき固有表現（正解固有表現）として含まれているかを列挙した正解リストとに基づきコンピュータ処理して生成する固有表現抽出規則生成システムであって、上記訓練用文書を上記記憶装置から読み出して単語分割し、各単語に品詞名と構成文字種情報を付加して、上記訓練用文書に含まれる固有表現を構成する単語列情報を生成して上記記憶装置に記録する品詞文字種付与手段と、上記単語列情報を上記記憶装置から読み出し、該読み出した単語列情報に含まれる単語が数字か固有名詞の場合もしくは当該単語列情報の末尾の単語が記号・単漢字・接尾語・接頭語・助詞を含む機能語のいずれでもない場合には該単語を変数化して、変数を含む単語列情報を求め、該変数を含む単語列情報と、上記記憶装置から読み出した上記正解リストとに基づき上記規則を生成する規則生成手段とを有することを特徴とする固有表現抽出規則生成システム。

2

【請求項4】　請求項1から請求項3のいずれかに記載の固有表現抽出規則生成システムにおいて、上記規則生成手段は、生成した規則に、該規則の生成に用いた上記固有表現が上記正解リスト中に現れるのべ回数を、該規則の優先度として付与する手段を有することを特徴とする固有表現抽出規則生成システム。

【請求項5】　請求項1から請求項4のいずれかに記載の固有表現抽出規則生成システムを具備し、該固有表現抽出規則生成システムにより生成された規則に基づきコンピュータ処理して任意の文書に含まれる固有表現を抽出する固有表現抽出装置であって、抽出した複数の候補固有表現に部分的な重なりがあれば、各候補固有表現の上記任意の文書における記載開始位置が早いものを優先して抽出し、上記記載開始位置が同じであれば記載終了位置が遅いものを優先して抽出する手段と、抽出した複数の候補固有表現が同じであれば、各候補固有表現の抽出に用いた各々の上記規則に予め付与された優先度の大きいものを優先して抽出する手段とを有することを特徴とする固有表現抽出装置。

【請求項6】　文書から固有表現を抽出するのに用いる規則を、予め記憶装置に記録された訓練用文書と、該訓練用文書の中のどの位置にどのような種類の固有表現が抽出されるべき固有表現（正解固有表現）として含まれているかを列挙した正解リストとに基づきコンピュータ処理して生成するシステムの固有表現抽出規則生成方法であって、上記訓練用文書を上記記憶装置から読み出して単語分割する第1のステップと、分割した各単語に品詞名と構成文字種を付加して、上記訓練用文書に含まれる固有表現を構成する単語列情報を生成する第2のステップと、上記正解リストの各正解固有表現を上記記憶装置から読み出して上記第2のステップで生成された各単語列情報と比較し、当該正解固有表現を抽出するための規則を生成して上記記憶装置に記録する第3のステップと、上記記憶装置から上記訓練用文書と上記規則とを読み出して各規則を上記訓練用文書に適用し、対応する固有表現（候補固有表現）を抽出して上記記憶装置に記録する第4のステップと、上記記憶装置から上記候補固有表現と上記正解リストの正解固有表現とを読み出して比較し、予め定められた算出手順に基づき、各候補固有表現の抽出に用いた各規則の適正度を算出する第5のステップと、該第5のステップで算出した適正度が予め定められた適正度より低い規則を上記記憶装置から削除する

第6のステップと、上記第5のステップで算出した適正度が予め定められた適正度範囲の規則を、当該適正度が高くなるように修正して上記憶装置に記録する第7のステップとを有することを特徴とする固有表現抽出規則生成方法。

【請求項7】　請求項6に記載の固有表現抽出規則生成方法において、上記第3のステップは、上記記憶装置から読み出した単語列情報に含まれる単語が、数字か固有名詞の場合もしくは当該単語列情報の末尾の単語が記号・単漢字・接尾語・接頭語・助詞を含む機能語のいずれでもない場合には該単語を変数化して、変数を含む単語列情報を求めるステップと、該変数を含む単語列情報と、上記記憶装置から読み出した上記正解リストとに基づき上記規則を生成するステップとを有することを特徴とする固有表現抽出規則生成方法。

【請求項8】　請求項6、もしくは、請求項7のいずれかに記載の固有表現抽出規則生成方法において、上記第4のステップは、上記解候補固有表現に、該解候補固有表現の上記訓練用文書における記載位置情報および該固有表現の抽出に用いた規則の識別情報を付与して上記記憶装置に記録するステップを有し、上記第5のステップは、上記記憶装置から上記解候補固有表現と上記正解リストを読み出して比較し、上記正解リストにある候補固有表現（正解候補固有表現）と、上記正解リストにないが該正解リストにある他の正解固有表現により出力が抑制された候補固有表現（中間候補固有表現）、および、上記正解リストになく且つ該正解リストにある他の正解固有表現によっても出力が抑制されない候補固有表現（不正解候補固有表現）に分類するステップと、各候補固有表現の抽出に用いた各規則毎に、該規則により抽出された上記正解候補固有表現の数と上記不正解候補固有表現の数を計数するステップとを有し、上記第6のステップでは、上記正解候補固有表現の数に対する上記不正解候補固有表現の数が予め定められた基準Ｔ１以上の規則を上記第4のステップで生成した規則群から削除し、上記第7のステップでは、上記正解候補固有表現の数に対する上記不正解候補固有表現の数が予め定められた基準Ｔ２以下の規則を、上記不正解候補固有表現の数が減少するよう修正することを特徴とする固有表現抽出規則生成方法。

【請求項9】　請求項6から請求項8のいずれかに記載の固有表現抽出規則生成方法において、上記第5のステップでは、上記記憶装置から同じ規則で抽出された複数の候補固有表現を読み出して、上記正解固有表現に一致する候補固有表現（正解候補固有表現）と一致しない候補固有表現（不正解候補固有表現）および上記正解固有表現に一致しないが他の上記正解候補固有表現により出力が抑制される候補固有表現（中間候補固有表現）に分け、上記正解候補固有表現と上記不正解候補固有表現のそれぞれの数に基づき上記適正度を算出し、上記第7のステップでは、上記適正度が予め定められた適正度範囲の規則（元の規則）を上記訓練用文書に適用して抽出された各候補固有表現のそれぞれの上記訓練用文書における前後の単語や該単語の文字種や品詞を求め、該前後の単語や該単語の文字種や品詞に基づき、上記各候補固有表現に含まれる上記不正解固有表現を抽出させない制約条件を生成して上記元の規則に加えることを特徴とする固有表現抽出規則生成方法。

【請求項10】　文書から固有表現を抽出するのに用いる規則を、予め記憶装置に記録された訓練用文書と、該訓練用文書の中のどの位置にどのような種類の固有表現が抽出されるべき固有表現（正解固有表現）として含まれているかを列挙した正解リストとに基づきコンピュータ処理して生成するシステムの固有表現抽出規則生成方法であって、上記訓練用文書を上記記憶装置から読み出して単語分割する第1のステップと、各単語に品詞名と構成文字種情報を付加して、上記訓練用文書に含まれる固有表現を構成する単語列情報を生成して上記記憶装置に記録する第2のステップと、上記記憶装置から読み出した単語列情報に含まれる単語が、数字か固有名詞の場合もしくは当該単語列情報の末尾の単語が記号・単漢字・接尾語・接頭語・助詞を含む機能語のいずれでもない場合には該単語を変数化して、変数を含む単語列情報を求め、該変数を含む単語列情報と、上記記憶装置から読み出した上記正解リストとに基づき上記規則を生成する第3のステップとを有することを特徴とする固有表現抽出規則生成方法。

【請求項11】　請求項6から請求項10のいずれかに記載の固有表現抽出規則生成方法において、上記第3のステップは、生成した規則に、該規則の生成に用いた上記固有表現が上記正解リスト中に現れるべき回数を、該規則の優先度として付与するステップを有することを特徴とする固有表現抽出規則生成方法。

【請求項12】　文書から固有表現を抽出するのに用いる規則を、予め記憶装置に記録された訓練用文書と、該訓練用文書の中のどの位置にどのような種類の固有表現が抽出されるべき固有表現（正解固有表現）として含まれているかを列挙した正解リストとに基づきコンピュータ処理して生成する方法の処理手順を既述したプログラムをコンピュータに読取り可能に記録する記録媒体であって、請求項6から請求項11のいずれかに記載の固有表現抽出規則生成方法における各ステップを、上記コンピュータに実行させるための処理プログラムを記録したことを特徴とする記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、コンピュータを用いて、文書に含まれる固有表現を抽出する技術に係わり、特に、固有表現を抽出するために用いる規則を効率的に生成するのに好適な固有表現抽出規則生成システム

と方法およびその処理プログラムを記録した記録媒体ならびに固有表現抽出装置に関するものである。

【0002】

【従来の技術】膨大な量の文書に含まれる情報についての質問に答えたり、文書を要約したり、データベース化したり、視覚化したりするためには、その文書から、人名や地名や組織名や日時などの固有表現を抽出する必要がある。この場合、コンピュータを利用して、予め各固有表現を登録した辞書を用意しておき、この辞書を検索することにより、文書からの固有表現の抽出を行うことができる。

【0003】しかし、実際の文書には、予め用意した辞書に含まれない新しい言葉が必ず存在するので、辞書の検索だけでは、正確な抽出結果は得られない。このような問題に対処するために、固有表現そのものと、その前後に含まれる単語の並びの出現パターンを規則化して得た多数の規則を予め人手により作成し、その規則に基づきコンピュータ処理して、対象の文書から、固有表現を抽出するという技術がある。

【0004】しかし、この技術では、規則同士が競合したり相互作用したりするため、それぞれの規則が意図したとおりに働くとは限らないので、作成された規則を、予め用意された訓練データに適用して、その結果に基づき、間違ったところを見つけ出して、規則を修正するという作業を何度も繰り返さなければならない。

【0005】ところが、ある規則を修正した結果、それまで正常に働いていた規則が影響を受けて、間違った答を出すようになることが少なくない。そのため、多数の規則の全てを意図したとおりに働くようにするためには、膨大な時間と労力を要する。

【0006】このような固有表現を抽出する規則をコンピュータを用いて自動的に生成する技術においても、規則の間の競合や相互作用のため、自動生成された規則同士をどのように組み合わせれば良い成績が得られるかは、組み合わせた規則（ルール）を再度、実際の文書に適用して、その結果を正解と比較して採点し、その結果に基づき、より良い成績が得られるように規則を追加したり削除したりする試行錯誤を繰り返すしかなく、多大な計算時間が必要である。

【0007】

【発明が解決しようとする課題】解決しようとする問題点は、従来の技術では、文書に含まれる固有表現を高い精度で抽出するための規則を生成できない点と、より良い規則（固有表現抽出規則）を生成するためには、各規則の組合せを修正する度に、実際の文書に適用し、その結果を正解と比較して採点し、各規則の組合せの試行錯誤を行うので、多大な計算時間が必要となってしまう点である。

【0008】本発明の目的は、これら従来技術の課題を解決し、高精度な固有表現抽出規則の生成を容易に短時間で行うことを可能とし、膨大な文書から所望の固有表現を正確に抽出することを可能となる固有表現抽出規則生成システムと方法およびその処理プログラムを記録した記録媒体ならびに固有表現抽出装置を提供することである。

【0009】

【課題を解決するための手段】上記目的を達成するため、本発明の固有表現抽出規則生成システムと方法では、まず、予め用意された訓練用文書を形態素解析して単語に分割し、品詞名や構成文字種などの情報を各単語に付加する。こうして得られた単語列から、固有表現を構成する単語列を取り出し、訓練用文書に対応して予め用意された正解リストを参照して経験則や最小汎化などの一般化手段によって多数の固有表現抽出用の規則（ルール）を生成する。そして、これらの規則をそれぞれ独立に訓練用文書に適用して、その規則が、訓練用文書のどの位置にマッチしたかの記録を記憶しておく。この記録に入っているものは、訓練用文書に対してシステムが出力する固有表現の候補となる。そして、複数の規則を組み合わせる場合には、それらの規則に対応する記録に入っている全ての候補の中から、競合関係と優先順位を考慮して、最終的に出力する候補の列を一定の明快な基準で選び出す。この結果、訓練用文書における不正解の頻度あるいは割合が非常に多い規則があれば、それを削除する。ただし、その規則が訓練用文書のどの位置で正解し、どの位置で不正解になっているかがわかる。そこで、正解の箇所の前後の単語列と、不正解の箇所の前後の単語列を比較して制約を加えることによって、訓練用文書における成績が良くなる規則が作れるかどうか判断できるので、成績が良くなる場合は制約を加えた規則を加える。

【0010】さらに、本発明の固有表現抽出装置では、上述の固有表現抽出規則生成システムを具備し、この固有表現抽出規則生成システムで生成された規則に基づき任意の文書中の固有表現を抽出すると共に、抽出した複数の固有表現に部分的な重なりがあれば、文書における記載開始位置が早いものを優先して抽出し、また、記載開始位置が同じであれば記載終了位置が遅いものを優先して抽出し、さらに、表現は同じであるが種類の異なる固有表現があれば、各固有表現の抽出に用いた各々の規則に予め付与された優先度の大きいものを優先して抽出する。

【0011】

【発明の実施の形態】以下、本発明の実施の形態を、図面により詳細に説明する。

【0012】図1は、本発明に係る固有表現抽出規則生成システムおよびそれを設けた固有表現抽出装置の構成例を示すブロック図であり、図2は、図1における固有表現抽出規則生成システムおよび固有表現抽出装置のハードウェア構成例を示すブロック図である。

7

【0013】図2において、21はCRT (Cathode Ray Tube)やLCD(Liquid CrystalDisplay)等からなる表示装置、22はキーボードやマウス等からなる入力装置、23はHDD (Hard Disk Drive) 等からなる外部記憶装置、24はCPU (Central Processing Unit) 24aや主メモリ24b等を具備して蓄積プログラム方式によるコンピュータ処理を行なう情報処理装置、25は本発明に係わるプログラムやデータを記録したCD-ROM (Compact Disc-Read Only Memory) もしくはDVD (Digital Video Disc/Digital Versatile Disc) 等からなる光ディスク、26は光ディスク25に記録されたプログラムおよびデータを読み出すための駆動装置、27はLAN (Local Area Network) カードやモデム等からなる通信装置である。

【0014】光ディスク25に格納されたプログラムおよびデータを情報処理装置24により駆動装置26を介して外部記憶装置23内にインストールした後、外部記憶装置23から主メモリ24bに読み込みCPU24aで処理することにより、情報処理装置24内に図1に示す固有表現抽出規則生成システムおよびそれを具備した固有表現抽出装置が構成される。

【0015】図1の固有表現抽出装置においては、訓練用文書1と、正解リスト2、固有表現抽出規則群5、改良後固有表現抽出規則群5a、訓練用記録7、新規文書11、および、抽出された固有表現のリスト13のそれぞれは、図2における外部記憶装置23もしくは主メモリ24b等に格納され、また、形態素解析・品詞文字種付与器3と、規則生成部4、訓練用規則適用部6、規則評価部8、規則削除部9、規則精緻部10、実施用規則適用部12のそれぞれは、図2における CD-ROM25に格納されたプログラムに基づき情報処理装置24内に構成される。

【0016】そして、形態素解析・品詞文字種付与器3と、規則生成部4、訓練用規則適用部6、規則評価部8、規則削除部9、規則精緻部10のそれぞれが本発明に係わる固有表現抽出規則生成システムを構成している。

【0017】形態素解析・品詞文字種付与部3は、訓練用文書1を単語分割して、各単語にその品詞名や構成文字種の情報を付加する。

【0018】規則生成部4は、形態素解析・品詞文字種付与部3の処理で得られる単語列を正解リスト2で与えられる抽出すべき固有表現のデータと突き合わせて、各固有表現を構成する単語列を取り出し、これを一般化して規則を生成する。その結果が固有表現抽出規則群5として図2における外部記憶装置23に記録される。

【0019】訓練用規則適用部6は、規則生成部4の生成結果で得られる固有表現抽出規則群5を訓練用文書1に適用する。その結果は訓練用記録7として図2における外部記憶装置23に記録される。

8

【0020】規則評価部8は、訓練用記録7に基づいて各規則を評価する。規則削除部9は、規則評価部8の評価結果に基づいて、成績の悪い規則を削除する。規則精緻部10は、成績が良くなるように規則を精緻する。

【0021】実施用規則適用部12は、このようにして改良された固有表現抽出規則群5（改良後固有表現抽出規則群5a）を、実際の新規文書11に適用して固有表現リスト13を得る。

【0022】訓練用規則適用部6と実施用規則適用部12はいずれも、規則群を文書に適用して固有表現を抽出するものであり、その処理内容はほぼ同じであるため、単一の装置で両者を兼ねることも可能である。ただし、実施用規則適用部12は、訓練用記録7を残す必要がないが、最終的な候補の選択を行なう必要がある点が異なる。

【0023】まず、実施用規則適用部12の動作、すなわち、本例の固有表現抽出規則生成システムで生成・改良された固有表現抽出規則群5、改良後固有表現抽出規則群5aを用いた固有表現抽出装置としての動作を説明する。

【0024】実施用規則適用部12は、固有表現を抽出したい新規文書11に対して、改良後固有表現抽出規則群5aを適用して、文書中に含まれる固有表現を抽出して固有表現リスト13を出力する。

【0025】例えば、「田中太郎賞選考委員会では、・・・」という新規文書11があるとすると、この文書中の固有表現として、「田中」、「太郎」、「田中太郎」という人名の候補と、「田中太郎賞」という人工物名の候補、さらに、「田中太郎賞選考委員会」という組織名の候補が考えられるが、一般には、その内で一番長い「田中太郎賞選考委員会」だけが固有表現として抽出され出力されることが望まれる場合が多く、この場合、これと重なっている「田中」や「太郎」などの他の候補（固有表現）は出力されるべきでない。

【0026】このような候補間の関係は、重なりに起因する競合関係と、各候補の優先順位による抑制関係に還元することができる。つまり、「田中太郎賞選考委員会」は「田中」などの他の候補と重なっているがために競合し、長い「田中太郎賞選考委員会」の優先順位が高く、短い他の候補を抑制していると考えることができる。

【0027】本例においては、実施用規則適用部12は、この考え方に基づき、まず、全ての規則を文書に適用することで、全ての固有表現の候補の集合（「田中」や「太郎」、「田中太郎」、「田中太郎賞」、「田中太郎賞選考委員会」などを含む）を求める。次に、これらの候補の中で同じ固有表現（上の各候補においては「田中」）が最初に現れるものの内で一番長いもの（上の各候補においては「田中太郎賞選考委員会」）を出力する。

【0028】このようにして一つの候補が出力されると、この候補と競合している他の全ての候補（「田中」、「田中太郎」、「田中太郎賞」）を候補の集合から削除する。候補の集合が空になるまで、この作業を繰り返すことにより、固有表現のリスト１３が得られる。

【0029】ただし、このように長さだけに着目して、各々競合する各候補からの選択の判断を行うだけでは、同じ長さの複数の候補がある場合に判断に困る。例えば「ホワイトハウス」は、地名と考えられる場合と組織名と考えられる場合があるので、同じ「ホワイトハウス」という文字列を地名とする候補と、組織名とする候補とが考えられる。

【0030】そこで、この２つの候補の間に、抽出するための優先順位を設ける。例えば、その前後の単語を考慮して、「ホワイトハウスの近くの公園で・・・」であれば地名の可能性が高く、「ホワイトハウスによれば・・・」であれば、組織名の可能性が高い。また、例えば、その出現頻度を考慮して、訓練用文書１に「ホワイトハウス」が地名として出現しているのが１回で、組織名として出現しているのが２０回とすれば、組織名と判断した方が正解する可能性が高い。

【0031】本例では、改良後固有表現抽出規則群５ａにおける各規則には、これらの条件を加味した優先度が付与されている。

【0032】実施用規則適用部１２は、このような優先度と、前述の固有表現の長さとを組み合わせて、各候補の優先順位を計算する。この優先順位の設定としてはさまざまな変種が考えられるが、上記のように、開始位置が一番早いものの中で、さらに終了位置が一番遅いものの内、優先度が一番高いものを選ぶのが明快である。つまり、候補の優先関係については、以下のような定義を基本とする。

【0033】①候補Ａの開始位置が候補Ｂの開始位置より早い(数字として小さい)ならば、候補Ａの方が優先される。

②候補Ａの開始位置と候補Ｂの開始位置が同じであれば、終了位置が遅い(数字として大きい)候補が優先される。

③両候補の開始位置と終了位置が全く同じであれば、予め規則で与えられた優先度ｕの大きい候補が優先される。

【0034】本例の固有表現抽出規則生成システムでは、このような実施用規則適用部１２による処理を容易とする固有表現抽出規則群５および改良後固有表現抽出規則群５ａを生成する。以下、このような優先関係を加味した規則の生成処理に係る固有表現抽出規則生成システムを構成する各部の動作について説明する。

【0035】まず、形態素解析・品詞文字種付与部３において、文書を単語列に分割する。典型的には形態素解析機能を有し、訓練用文書１や新規文書１１などの与えられた文書を単語分割して、各単語に品詞名とその単語を構成する文字の種類（構成文字種情報）を付与したデータ構造を作り、そのリストを作成する。

【0036】例えば、「東京製鉄の中野社長は・・・」という文があると、形態素解析により「東京」は固有名詞、「製鉄」は普通名詞、「の」は助詞、「中野」は固有名詞、「社長」は普通名詞、「は」は助詞、という結果が得られる。

【0037】また、「東京」は複数の漢字で構成されており、「の」はひらがなである。従って、形態素解析・品詞文字種付与部３は、この文に対して、例えば以下のようなデータ構造からなるリストを出力する。［(東京，複数漢字，固有名詞)、(製鉄，複数漢字，普通名詞)、(の，ひらがな，助詞)、・・・］

【0038】一方、正解リスト２は、訓練用文書１の中のどの位置にどのような種類の固有表現が含まれているかを列挙したものであり、「東京製鉄の中野社長は・・・」という訓練用文書１に対応して予め用意される正解リスト２は、例えば、次のようなデータからなる。

【0039】

0	3	東京製鉄	組織名
5	6	中野	人名
20	23	3月9日	日付
30	32	岡山県	地名

【0040】このリストにおいて、最初の行は、この文書の「0文字目から３文字目の位置」が「東京製鉄」という「組織名」をその種類とする固有表現であり、次の行は「5文字目から６文字目の位置」が「中野」という「人名」をその種類とする固有表現であることを示している。このように、本例の正解リスト２においては、各固有表現の開始位置と終了位置を示す数字の対で、当該固有表現の位置を略称する。

【0041】規則生成部４は、このような正解リスト２と、形態素解析・品詞文字種付与部３の出力する単語列とを突き合わせて、固有表現を変数化等して、例えば、次のような固有表現の抽出規則を生成する。

【0042】anytag(3) <-- <@(組織名，21)，word(，複数漢字，固有名詞)，word(製鉄，複数漢字，普通名詞)，>@(組織名)．

【0043】この規則（ルール）は、番号「２１」が付与された規則であり、任意の（変数化された）漢字の固有名詞があり（「word(，複数漢字，固有名詞)」）、その次の単語が「製鉄」という複数漢字の普通名詞であれば（「word(製鉄，複数漢字，普通名詞)」）、その２単語が、「組織名」の固有表現の候補として考えられるという意味の規則である。

【0044】このような規則（ルール）の生成は、より一般的には以下のように表せる。まず、固有表現は、Ｎ＋１単語［(w0，c0，p0)，・・・，(wi，ci，pi)，・・・，(wN，cN，pN)］でできているとす

る。ここでwiは単語（「製鉄」、「中野」など）、ciは構成文字種（「複数漢字」や「数字」など）、piは品詞名（「固有名詞」、「普通名詞」など）である。

【0045】実際には、前後の数つかの単語も、固有表現かどうかを判断するのに重要な手がかりとなるので、含めて考えるのが一般的であるが、ここでは単純化して、固有表現に含まれる単語だけを考える。

【0046】次に、このような単語列から、最小汎化などの既存の一般化技術を用いることによって、規則（ルール）を生成する。しかし、本例では、次のようにして簡単に生成する。

【0047】すなわち、訓練用文書１に含まれる固有表現を構成する具体的な単語列［(w0, c0, p0), ・・・, (wi, ci, pi), ・・・, (wn, cn, pn)］に、以下に述べる経験則を適用して、変数を含むリスト［(w0', c0', p0'), ・・・, (wi', ci', pi'), ・・・, (wn', cn', pn')］を得て、次のような規則を作る。

【0048】anytag(u) <— <@(t+df, k), word(w0', c0', p0'), ・・・, (wi', ci', pi'), ・・・, word(wn', cn', pn'), >@(t−dt).

【0049】ここで「t」は、固有表現の種類（例えば「組織名」）を表す。「+df」は、この固有表現の開始位置を何文字右にずらすかを表し、最初の単語の文字数未満の非負整数である。また、「−dt」は固有表現の終了位置を何文字左にずらすかを表し、最後の単語の文字数未満の非負整数である。

【0050】例えば、「厚木市内で・・・」という訓練用文書１があり、正解リスト２によればこの内の「厚木市」が地名であるにもかかわらず、形態素解析・品詞文字種付与部３の形態素解析で、「厚木」、「市内」、「で」というように単語分割された場合、固有表現を構成する単語列は、［(厚木, 複数漢字, 固有名詞), (市内, 複数漢字, 普通名詞)］となり、最後の１文字（「内」）が余分である。そこで終了位置を一文字左にずらすために、「dt=1」とする。尚、開始位置はずらさないので、「df=0」である。

【0051】また、上述の規則（ルール）における「k」は、この規則につけられた番号であり、「u」はこの規則の優先度である。

【0052】各変数を含むデータ(wi', ci', pi')は、訓練用文書１に含まれる具体的な固有表現に対応するデータ(wi, ci, pi)に対して、以下の経験則を、上から順に調べ、最初に当てはまったものを適用することによって得る。

【0053】①「i」が「0」か「N」で、固有表現の境界を含む場合（df>0またはdt>0）は、これらを変数化しない。規則（ルール）の「df」と「dt」は、元になった固有表現に対する値をそのまま利用する。

②数字の場合は「wi」を変数化する。

③固有名詞の場合は「wi」を変数化する。

④リストの最後の単語か、記号・単漢字・接尾語・接頭語・助詞などの機能語であれば、変数化しない。

⑤それ以外であれば「wi」を変数化する。

【0054】各固有表現に対して以上の処理を適用することにより、固有表現抽出規則群５を自動的に生成することができる。

【0055】また、各規則の優先度（u）としては、例えば、その規則の元になった固有表現が正解リスト中に現れる「のべ回数」を採用する。これにより、正解回数の少ない規則（前述の例では、地名としての「ホワイトハウス」）が正解回数の多い規則（組織名としての「ホワイトハウス」）を正当な理由もなく抑制してしまうことが避けられる。

【0056】こうして規則生成部４により得られた各規則（固有表現抽出規則群５）を、訓練規則適用部６において、訓練用文書１の単語列に適用することにより訓練用記録７を得る。すなわち、訓練規則適用部６では、訓練用文書１の先頭から末尾まで、各規則がマッチする位置を順に調べていき、マッチしたら、それを候補として訓練用記録７に追加することを繰り返す。

【0057】訓練用記録７には、具体的には、後で各候補間の競合関係や抑制関係の比較をして、最終的な出力ができるように、ルール番号（k）や、マッチした位置、固有表現の種類（t）などのデータを記録しておく。

【0058】このような訓練規則適用部６による処理を、固有表現抽出規則群５の全ての規則に対して行ない、訓練用記録７を作り出す。

【0059】尚、ボトムアップ型の構文解析を用いれば、複数の規則の適用結果を効率良く一度に得ることも可能である。

【0060】規則評価部８は、このようにして作成された訓練用記録７を読み出して、各規則の成績を採点する。採点の仕方としては様々な基準を用いることができるが、不正解になった回数や割合による評価を用いれば簡単である。しかし、各規則の不正解回数は、厳密には、どのような規則と組み合わせて用いるかに依存するため、どの規則を採用するか未定のこの時点では、正確な数字を得られない。そこで、各規則（R）の記録を以下のように分類して考える。

【0061】

（○）規則Rの元になった固有表現とマッチして得られた候補、つまり、他の候補に抑制されなければ必然的に正解になるもの（正解候補固有表現）。

（△）競合する別の固有表現が正解リスト２に登録されており、それに抑制されるもの。つまり、その固有表現が正解になれば出力が抑制されるので、精度の高い規則群においては、成績を下げない可能性の高いもの（中間

候補固有表現）。

（×）それ以外のもの、つまり、抑制する正解固有表現がないため、精度の高い規則群においては、間違った候補を出力して成績を下げる可能性が高いもの（不正解候補固有表現）。

【0062】規則評価部8は、各規則に対して「○」、「△」、「×」の回数を数え、この「×」の回数を不正解の回数、「○」の回数を正解の回数の代用として採用する。尚、単純に「△」を全て不正解と考えると、「田中」のように短い固有表現を抽出する規則が不利になるので避けた方が良い。そのため、規則評価部8では、以下のような方法で不正解回数を数える。

【0063】すなわち、規則評価部8は、訓練用記録7を前から順に読み、規則Rが訓練用文書1の位置Lで適用されており、規則Rが付与する固有表現のタイプ（地名や人名などの区別）がTであり、そのタイプTと位置Lの対が正解リスト2に正解として含まれておらず、さらに、位置Lに重なる位置に正解の固有表現が存在しないか、存在しても、その正解に対応する候補より規則Rによる候補の方が優先順位において優位であれば、規則Rの不正解回数を1増やす。これを訓練用記録7の終わりに達するまで繰り返す。

【0064】規則評価部8が、各規則の「○」、「△」、「×」の個数を数えると、この結果を参照して、規則削除部9と規則精錬部10が固有表現抽出規則群5に修正を加える。

【0065】規則削除部9は、固有表現抽出規則群5の規則の内、例えば、「×」の個数が「○」の個数を超える規則を削除する。規則精錬部10は、固有表現抽出規則群5の規則の内、例えば、成績が「×」の個数が「○」の個数の半分以上ある規則に、前後の単語などに関する制約情報を加えて、当該規則の成績が より良くなるようにする。

【0066】例えば、固有表現の前後2単語ずつを含めて考えると、上記規則で抽出され、「○」や「×」に評価されて分類された各固有表現のそれぞれにおいて、[(w-2, c-2, p-2), (w-1, c-1, p-1), (w0, c0, p0), ・・・, (wn+1, cn+1, pn+1), (wn+2, cn+2, pn+2),]という単語リストが各々に考えられる。そこで、各固有表現毎に(w-2, c-2, p-2, w-1, c-1, p-1, wn+1, cn+1, pn+1, wn+2, cn+2, pn+2)という特徴のリストを考え、「○」に分類された固有表現の場合を正例、「×」に分類された固有表現の場合を負例と考えれば、これは典型的な帰納学習の課題であり、既存の機械学習の手法がそのまま利用できる。

【0067】例えば、決定木による学習を用いることにより、前後の幾つかの単語の内、どの単語のどの性質の値を残し、他を変数化すべきかが決定できる。具体例として、「×」に分類された固有表現が「10」個抽出され、その内、「8」個の固有表現において、その前の単語(w-1)として「w1」が特定されれば、以下のようにして元の規則に制約条件｛w-1'≠ w1｝を加え、前の単語(w-1)として「w1」を有する固有表現が抽出されないように制約する。

【0068】anytag(u) <— word(w-1', c-1', p-1'), <@(t+d f, k), word(w0', c0', p0'), ・・・, (w1', c1', p1'), ・・・, word(wn', cn', pn'), >@(t−d t), ｛w-1'≠ w1｝。

【0069】こうして得られた規則は、元の規則より制約が強いので、元の規則がマッチした部分と同じところにしかマッチしない。従って、訓練用文書1全体に適用しなくても、訓練用記録7に残っている元の規則のマッチした部分にのみ適用すれば、新しい規則の成績はわかる。

【0070】このように本例では、規則の改良が、他の規則とほぼ独立に行なえる。以上によって、元の規則（固有表現抽出規則群5）から、より成績の良い規則（改良後固有表現抽出規則群5ａ）を生成する。

【0071】図3は、本発明に係わる固有表現抽出規則生成方法の処理手順例を示すフローチャートである。

【0072】本例は、図1における固有表現抽出規則生成システムにおける形態素解析・品詞文字種付与部3、規則生成部4、訓練用規則適用部6、規則評価部8の各処理動作を示すものであり、まず、形態素解析・品詞文字種付与部3において、訓練用文書1を形態素解析して単語に分割し（ステップ３０１）、各単語に品詞と文字種などの情報を付加する（ステップ３０２）。

【0073】次に、規則生成部4において、正解リスト2の固有表現と、その近傍にある単語からなる単語列を抜き出して（ステップ３０３）、正解単語列に経験則等を適用して、抽出規則を生成し（ステップ３０４）、固有表現抽出規則群5として記録する。

【0074】そして、訓練用規則適用部6において、このようにして生成した抽出規則を、訓練用文書1に適用して、その結果得られた固有表現を候補として記録する（ステップ３０５）。

【0075】さらに、規則評価装置8において、各抽出規則で抽出された固有表現の正解度（○、△、×）を求めて分類し、それに基づき、各抽出規則の適正度を採点する（ステップ３０６）。

【0076】その採点の結果、修正不可能な成績の悪い（適正度の低い）規則群は、規則削除部9において削除し（ステップ３０７）、また、修正により適正度が高まる規則群には、規則精錬部10において当該修正を加えて、新規則とし（ステップ３０８）、改良後固有表現抽出規則群5ａに記録する。ステップ３０５からの処理を繰り返すことにより、より成績の良い規則群の生成が可能となる。

【0077】図4は、図1における固有表現抽出装置の

処理動作例を示すフローチャートである。本例は、図1に示す固有表現抽出装置における、新規文書11に対する処理動作を示すものであり、まず、形態素解析・品詞文字種付与部3において、新規文書11を形態素解析して単語に分割し（ステップ401）、各単語リストに品詞と文字種などの情報を付加する（ステップ402）。

【0078】次に、実施用規則適用部12において、各単語リストに、改良後固有表現抽出規則群5aの各抽出規則を適用して、各固有表現を候補としてリストアップし（ステップ403）、全ての候補に対して以下の優先制御処理を行う（ステップ404）。すなわち、各候補の中で最優先の候補を出力し（ステップ405）、この出力された候補と競合する候補を削除する（ステップ406）。

【0079】以上、図1～図4を用いて説明したように、本例の固有表現抽出規則生成システムと方法では、まず、予め用意された訓練用文書1を形態素解析して単語に分割し、品詞名や構成文字種などの情報を各単語に付加し、こうして得られた単語から、固有表現を構成する単語列を取り出し、予め訓練用文書1に対応して用意された正解リスト2を参照して経験則や最小汎化などの一般化手段によって多数の固有表現抽出規則を生成する。

【0080】次に、これらの抽出規則をそれぞれ独立に訓練用文書1に適用して、その規則が訓練用文書1のどの位置にマッチしたかの記録を用意しておく。この記録に入っているものは、訓練用文書1に対してシステムが出力する固有表現の候補となる。

【0081】そして、複数のルールを組み合わせる場合には、それらのルールに対応する記録に入っている全ての候補の中から、競合関係と優先順位を考慮して、最終的に出力する候補の列を一定の明快な基準で選び出す。この結果、訓練用文書1における不正解の頻度あるいは割合が非常に多い規則があれば、それを削除する。ただし、その規則が訓練用文書のどの位置で正解し、どの位置で不正解になっているかがわかる。そこで、正解の箇所の前後の単語列と、不正解の箇所の前後の単語列を比較して制約を加えることによって、訓練用文書における成績が良くなる規則が作れるかどうか判断できるので、成績が良くなる場合は制約を加えた規則を加える。

【0082】このように、本例によれば、固有表現を含む訓練用文書と、その文書の中のどの位置にどのような種類の固有表現が含まれているかを列挙した正解リストを与えると、システムがこの正解に基づいて固有表現抽出規則を生成するので、人間が多大な労力を払って抽出規則を書き下す必要がなくなる。

【0083】さらに、予め用意された訓練用文書1に対して出力される個々の規則の評価を求め、次に、複数の規則を種々に組み合わせた場合の各評価値を、個々の規則の評価値から簡単に計算できる。これによって、良い成績が得られる規則の組み合わせを求める際の試行錯誤に要する処理時間を短縮することができる。また、このような固有表現抽出規則の改良が、他の規則とほぼ独立して行なえるため、精度を向上させることが容易になる。

【0084】また、本例の固有表現抽出装置では、訓練用文書と正解リストに基づいて生成され、かつ、改良された規則を新規文書11に適用して、この新規文書11から固有表現を自動的に抽出すると共に、抽出した複数の固有表現に部分的な重なりがあれば、文書における記載開始位置が早いものを優先して抽出し、また、記載開始位置が同じであれば記載終了位置が遅いものを優先して抽出し、さらに、表現は同じであるが種類の異なる固有表現があれば、各固有表現の抽出に用いた各々の規則に予め付与された優先度の大きいものを優先して抽出するので、適切な固有表現のみに限定された抽出が可能である

【0085】尚、本発明は、図1～図4を用いて説明した例に限定されるものではなく、その要旨を逸脱しない範囲において種々変更可能である。例えば、本例では、規則に制約を付加する際、候補固有表現の訓練用文書における前後の単語に基づき制約を設けているが、当該単語の文字種（漢字、カタカナ、・・・）や品詞（名詞、動詞、・・・）等に関して制約を設けることでも良い。

【0086】また、本例では、光ディスク25を記録媒体として用いているが、FDを記録媒体として用いることでも良い。また、プログラムのインストールに関しても、通信装置27を介してネットワーク経由でプログラムをダウンロードしてインストールすることでも良い。

【0087】

【発明の効果】本発明によれば、予め用意された訓練用文書と、その文書の中のどの位置にどのような種類の固有表現が含まれているかを列挙した正解リストとに基づき、固有表現を抽出するための規則を自動生成するので、人間が多大な労力を払って抽出規則を書き下す必要がなくなり、さらに、自動生成した規則を訓練用文書に適用してその結果と正解リストとを比較し、各規則毎の適正度を求め、この各規則毎の適正度に基づき、各規則を組み合わせた場合の適正度を求めることができるので、固有表現抽出規則の改良が、他の規則とほぼ独立して行なうことができ、精度を向上させることが容易になり、高性能な固有表現抽出装置を容易に実現することが可能である。

【図面の簡単な説明】

【図1】本発明に係る固有表現抽出規則生成システムおよびそれを設けた固有表現抽出装置の構成例を示すブロック図である。

【図2】図1における固有表現抽出規則生成システムおよび固有表現抽出装置のハードウェア構成例を示すブロック図である。

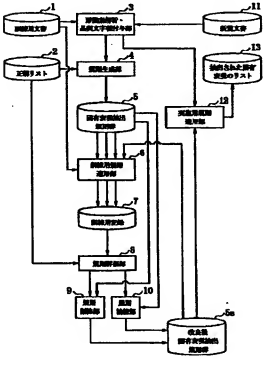
【図3】本発明に係わる固有表現抽出規則生成方法の処理手順例を示すフローチャートである。

【図4】図1における固有表現抽出装置の処理動作例を示すフローチャートである。

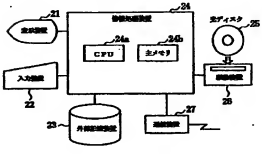
【符号の説明】

1：訓練用文書、2：正解リスト、3：形態素解析・品詞文字種付与部、4：規則生成部、5：固有表現抽出規則群、5a：改良後固有表現抽出規則群、6：訓練用規則適用部、7：訓練用記録、8：規則評価部、9：規則削除部、10：規則精練部、11：新規文書、12：実施用規則適用部、13：抽出された固有表現のリスト、21：表示装置、22：入力装置、23：外部記憶装置、24：情報処理装置、24a：CPU、24b：主メモリ、25：光ディスク、26：駆動装置、27：通信装置。

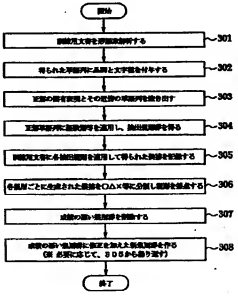
【図1】



【図2】



【図3】



【図4】